基于多层次注意力机制一维 DenseNet 的音频事件检测 *

杨吕祥，胡　燕†

(武汉理工大学 计算机科学与技术学院，武汉 430070)

摘　要：在音频事件检测任务中，目标音频易受背景噪声等因素的干扰，并且其在音频信号流中存在的比例不高，针对这些问题，提出一种多层次注意力机制一维 DenseNet(dense convolutional network)音频事件检测模型。首先，使用一维 DenseNet 模型进行帧级检测能有效地检测音频事件发生的开始和结束时间；其次，在一维 DenseNet 模型中引入多层次注意力机制，这使得不同模块的感知特性随着网络层数的加深而自适应地变化。因此，模型可以在不同的网络层次自动选择和关注重要的目标帧而抑制不相关的背景帧。在 DCASE 2017 任务 2 的开发数据集上的实验表明，该方法的整体性能较传统的深度学习方法有进一步提高。

关键词：音频事件检测；深度学习；DensetNet；多层次注意力机制

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.11.0867

Sound event detection based on ld DenseNet with multi-level attention

Yang Lyuxiang, Hu Yan†

(*School of Computer Science & Technology*, *Wuhan University of Technology*, *Wuhan* 430070, *China*)

Abstract: In sound event detection tasks, the target event was susceptible to background noise, and was not present in a significantly high portion of time frames of each signal. To solve the problem, this paper proposed a new method of sound event detection based on one-dimensional Dense Convolutional Network(DenseNet) with multi-level attention mechanism. Firstly, it used the one-dimensional DenseNet for frame-wise detection, which was effective in finding the precise onset and offset time. Then, it embedded the multi-level attention mechanism in the one-dimensional DenseNet model, which made the attention-aware features from different modules change adaptively as layers went deeper. Therefore, the model could automatically select and attend on important frames for the targets while ignoring the unrelated parts (e. g. , the background noise segments) . Finally, this work evaluated the model using DCASE 2017 Task 2 development dataset. Results show that the overall performance of the method has further improved than the conventional deep learning method.

Key words: sound event detection; deep learning; DenseNet; multi-level attention mechanism

0　引言

音频事件检测(sound event detection, SED) [1]是计算机听觉场景分析(computational auditory scene analysis, CASA) [2]领域的一种特定任务，它根据音频流的声学内容对音频中的事件进行分类和定位，目的是为每个被检测到的事件分配一个类标签以及确定音频事件发生的起始和结束时间，进而达到感知和理解周围环境的目的。音频事件检测的应用场景十分广泛，包括鸟声检测[3]、音频监控[4]以及多媒体事件检测[5]等，因此，音频事件检测相关的研究受到越来越多的研究机构以及学者的青睐。例如，音频事件检测与分类比赛(detection and classification of acoustic scenes and events, DCASE[1])自 2013 年举办以来吸引了大量的参与者。

目前，真实环境下的音频事件检测相关的研究有很多，传统的机器学习方法如隐马尔可夫模型[6]、非负矩阵分解[7]以及随机森林[8]等。但在传统的机器学习方法中，往往需要复杂的特征工程提取音频信号特征。为了解决以上缺点，有学者引入深度学习来进行音频事件检测。文献[9]提出一种深度卷积神经网络(convolutional neural networks，CNN)的端到端的学习框架自动从音频样本数据中学习特征，克服了传统机器学习需要手工提取音频特征的缺点。Cakir 等人[10]用多标签的深度神经网络（deep neural network, DNN）模型检测音频样本中重叠的音频事件。但是由于传统的深度学习模型随着网络层数越来越深，架构越来越复杂，会引起梯度消失问题。为了缓解音频事件检测中梯度消失的问题，Dang 等人[11]将目前最先进的图像分类网络模型 DensetNet[12]应用于音频事件检测。DensetNet 的核心思想是建立了卷积层的前层与后层之间的密集连接，即保证在网络中层与层之间最大程度的信息传输的前提下，直接将所有层连接起来，这不仅缓解了梯度消失问题，而且有利于提取音频信号更深层次的特征、增强了特征之间的传递并提高了系统的性能。

在音频事件检测中，为了解决背景噪声对目标事件的干扰，Phan 等人[13]改进了 DNN 中权重损失函数，将目标音频事件与背景音频事件分为前景类与背景类，通过权重函数中的惩罚因子抑制背景噪声，但需要人工调节权重损失函数的惩罚因子，增加了不必要的调参工作。随着基于注意力机制的神经网络被广泛应用于文本分类[14]以及情感分类[15, 16]任务，也有学者将注意力机制引入音频事件检测任务中。如徐等人[17]在音频标注任务中，在卷积门限循环神经网络(convolutional gated recurrent neural network, CGRNN)基础上引入了注意力机制增强目标音频事件的权重以及抑制不相关的背景音频噪声。Turab 等人[18]将最新的胶囊网络（capsules

收稿日期：2018-11-16；**修回日期**：2019-01-21　**基金项目**：湖北省自然科学基金重点类项目（2017CFA012）

作者简介：杨吕祥（1992-），男，湖北仙桃人，硕士，主要研究方向为音频信号检测；胡燕（1965-），女（通信作者），湖北武市人，教授，硕导，博士，主要研究方向为信息检索、数据挖掘、通信网络（huyan@whut.edu.cn）.

networks，CapsNet）模型与注意力机制结合学习音频信号最显著的特征来检测大规模弱标签音频事件，在 DCASE 的任务中取得了突破性的成绩。Kong 等人[19]提出将多示例学习模型与单层注意力机制结合在 Google 发布的大规模音频数据集上的性能超越了 Google 发布的基准系统。然而单层注意力模块忽略了中间神经网络的大量信息，因此，Lee 等人[20]通过连接不同中间神经网络层提高了分类器的性能。

受网络模型 DensetNet 以及注意力机制的启发，在一维 DenseNet 网络模型中引入了多层次注意力机制，这样既能提取音频信号更深层次的特征，又能自动选择和关注重要的目标音频帧。由此，本文提出了多层次注意机制一维稠密连接卷积神经网络（multi-level attention 1d densely connected convolutional networks，MLA-DCNNs）模型。首先，MLA-DCNNs 模型以一维对数梅尔频谱作为网络的输入特征，并采用一维 DenseNet[12]模型方法进行帧级检测，使得模型不仅能有效的检测音频事件发生的起始时间，而且有利于音频特征的重用，降低了网络结构的参数数量；其次，在一维 DenseNet 网络模型中引入了多层次注意力机制。多层次注意力机制由多个堆叠的注意力稠密模块和全局注意力机制组成，其中注意力稠密块是在 DensetNet 的稠密块中引入局部注意力机制，全局注意力机制跨越多个稠密块。因此，在 MLA-DCNNs 中不同模块的注意力感知特性随着网络层数的加深而自适应地变化，模型可以在不同的网络层次自动选择和关注重要的目标帧而抑制不相关的背景帧。在通用的 DCASE 2017 数据集上实验，并验证了 MLA-DCNNs 的有效性。

1 音频事件检测模型

为了关注更多有价值的音频信息，本文提出了多层次的注意力机制一维 DenseNet 模型 MLA-DCNNs 对音频事件进行检测。MLA-DCNNs 模型主要是由四个主要的模块组成：a)音频特征的提取模块；b)一维的 DensetNet 网络模块；c)多层次注意力模块，即一维 DenseNet 模型中引入局部注意力机制模块和全局注意力机制模块；d)分类模块。MLA-DCNNs 模型结构如图 1 所示。

1.1 音频特征

在音频事件检测中，对数梅尔谱特征[10, 18]作为音频特征的深度神经网络取得很好的效果。因此，本文用对数梅尔谱特征作为网络模型的输入特征。为了提取对数梅尔谱，首先，将音频样本进行加窗、分帧，其中每帧帧长设置为 40 ms，帧移设置为 20 ms；其次，通过短时傅里叶变换将时域信号转换为频域信号得到频谱；然后，将每一帧的频谱通过 128 个 Mel 滤波器组得到梅尔频谱；接着，将梅尔频谱在幅值方向取对数得到一维的对数梅尔谱特征；最后，对训练数据集的样本特征进行归一化处理，其中均值为 0，标准差为 1。

1.2 一维 DenseNet 网络层

SED 任务中，很多研究者利用类似图像分析的方式使用二维的音频特征，如声谱图、MFCC[10]、梅尔频谱[17]等作为卷积神经网络模型的输入特征，部分学者使用了三维音频特征[13]提取谱图中有意义的频谱信息以及时间位置信息。然而二维的 CNNs 模型方法分析的是块级而不是帧级音频特征，在音频事件检测中需要准确的预测事件发生的时间位置，因此，相比二维频谱特征，一维的频谱特征更有能有效地检测音频事件发生的时间位置。

为了有效的检测音频事件发生的时间位置以及更充分地利用帧级音频特征，本文采用一维的 DensetNet 模型。相比传统的卷积神经网络，DenseNet 模型建立了卷积神经网络的前层与后层之间的密集连接，即在保证网络层与层之间最大程度的信息传输的前提下，直接将所有的层连接起来。在传统的卷积神经网络中，如果网络有 L 层，那么就会有 L 个连接。但是在 DenseNet 中，L 层网络有 $L*(L+1)/2$ 个连接，简言之，每一层的输入来自前面所有层的输出。例如，第 0 层到 $L-1$ 层的输出特征图通道数目分别为 $x_0, x_1, \cdots, x_{l-1}$，则第 L 层计算方法如下：

$$x_l = H_L([x_0, x_1, \cdots, x_{l-1}]) \tag{1}$$

其中：$[x_0, x_1, \cdots, x_{l-1}]$ 表示将 0 层到 $l-1$ 层的输出特征图做通道的合并；$H_L(\bullet)$ 代表三种操作的组合函数，分别是 batch normalization(BN)[21]、rectified linear unit(ReLU)[22]、卷积操作。

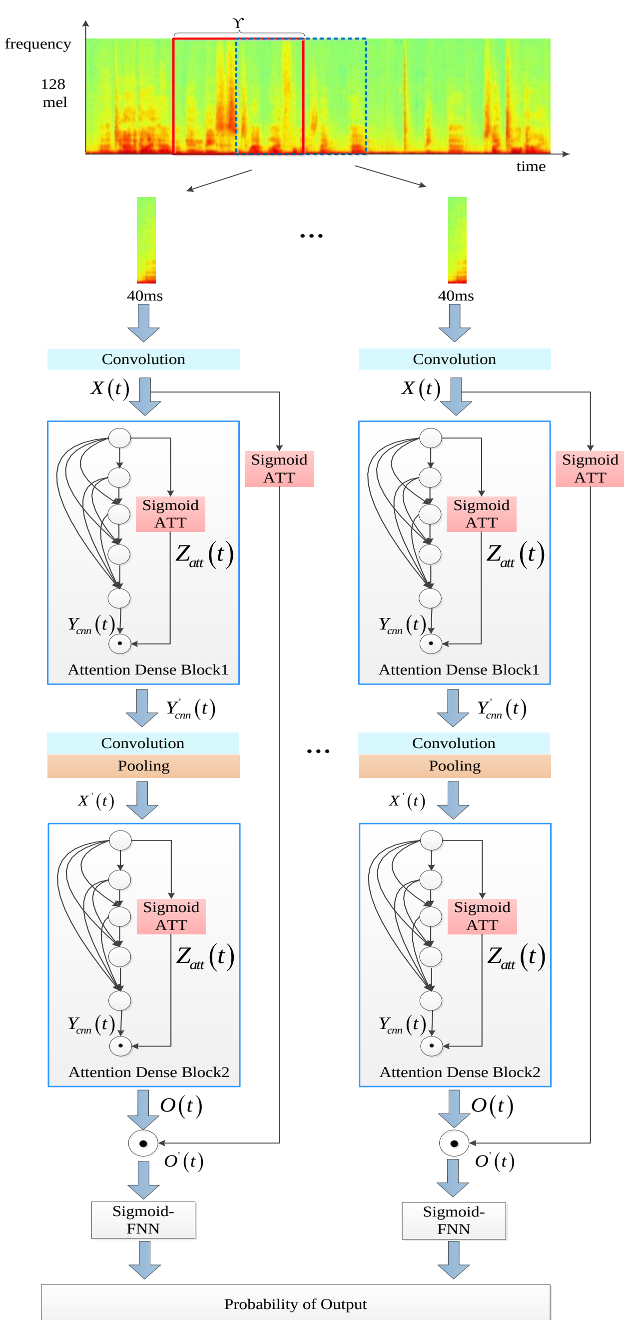


图 1 多层次注意力机制一维 DenseNet 音频事件检测模型

Fig. 1 Framework of proposed ld densenet with multi-level attention for sound event detection

正是由于 DenseNet 网络模型采用这种密集连接方式，使得模型具有以下优点：a)缓解了梯度消失；b)增强了特征的

传递；c)更有效地利用了特征；d)减少了网络参数数量。

因此，本文提出了用一维 DenseNet 网络模型提取音频帧级特征。一维 DenseNet 模型主要由若干个稠密块(dense block)和连接两个连续稠密块的转换层(transition layer)组成。每个稠密块由若干个连续的 BN[21]、ReLU[22]、1×1 以及 3×1 卷积层组成。其中 BN[21]是为了降低样本间的差异，其对每一层的输入作标准化处理，使样本输入的均值趋近于 0，标准差趋近于 1。ReLU[22]为网络的激活函数。稠密块中每层输出的特征图需要在通道的维度上进行合并，因此每个稠密块中的特征图大小需保持一致。在稠密块中，后面层的输入来自前面所有层的输出，因此即使稠密块中每层输出的特征图数目 k 很小，但是做通道连接之后，第 k 层的通道数 $k_0+k\times(k-1)$ 仍然很大，其中 k_0 表示输入层的特征图数目。为了减少输入特征图的数量以及融合各个通道的特征，在稠密块中采用 1×1 卷积层。转换层由 BN[21]层、1×1 卷积层以及 2×1 的平均池化层组成。为了进一步提高模型的紧凑型，减少转换层中的特征图数量，在转换层采用 1×1 的卷积层降低特征图的数目，例如，一个稠密块中包含 m 个特征图，经 1×1 的卷积层后生成 $\lfloor\theta m\rfloor$ 输出特征，其中 $0<\theta\le 1$ 称为压缩因子。转换层中的池化层降低了特征图的尺寸。

1.3　多层次注意力机制模块

多层次注意力机制模型由不同的模块堆叠而成，不同模块的注意力感知特性随着层数加深而自适应地变化，因而模型可以自动选择和关注重要的目标帧而忽略不相关的背景帧。而全局注意力模块为了获取全局显著特征。

1.3.1 注意力稠密模块

注意力稠密模块(attention dense block，ATT-DB)由一维稠密块以及 Sigmoid 层组成，如图 2 所示。由于每个注意力稠密模块结构相似，所以主要分析第一个模块。第 t 帧的注意系数 Z_{ATT1} 表示当前帧的重要程度，则有 Z_{ATT1} 表达式为

$$Z_{ATT1}(t)=\delta(W_{ATT1}*X(t)+b_{ATT1}) \tag{2}$$

其中：$X(t)$ 表示第 t 帧的输入特征；δ 表示 Sigmoid 函数；W_{ATT1} 表示 ATT-DB 的权值向量以及表示 ATT-DB 偏移量；Z_{ATT1} 表示第 t 帧注意力因子，因此只由全连接层与 Sigmoid 层组成。将预测的注意力因子与一维稠密块的输出做元素积运算来抑制背景噪声，计算公式如下：

$$Y'_{ADB1}(t)=Z_{ATT1}(t)*Y_{ADB1}(t) \tag{3}$$

其中：Y_{ADB1} 表示一维稠密连接模块输出；Y'_{ADB1} 表示加权后的输出特征。这种注意力机制加权过程可以选择重要的目标音频帧，同时抑制不相关的帧。

局部注意力模块是由多个 ATT-DB 堆叠而成。在真实场景下的音频事件检测中，堆叠的注意力模块结构可以逐渐细化特征，使不同模块的注意力感知特性随着层数加深而自适应地变化，因而模型可以在不同的层次自动选择和关注重要的目标帧而抑制不相关的背景噪声。

1.3.2 全局注意力机制

堆叠网络结构可以使不同层次的注意力机制关注不同层次的音频特征，而全局注意力机制是为了获取全局的显著特征。全局注意力因子跨越多个注意力稠密模块与最后一个注意力稠密模块输出的特征图做元素积运算。全局注意力机制由全连接层与 Sigmoid 层组成，如图 2 所示。

全局注意力因子计算公式与 ATT-DB 模块中注意力因子公式类似，全局注意力因子 Z_{GLB} 计算公式如下：

$$Z_{GLB}(t)=\delta(W_{GLB}*X(t)+b_{GLB}) \tag{4}$$

其中：δ 表示 Sigmoid 函数；W_{GLB} 表示全局注意力机制的权值向量；b_{GLB} 表示全局注意力机制的偏移量；Z_{GLB} 表示第 t 帧的全局注意力因子。在音频事件检测中，如果存在目标音频事件，则全局注意力因子 Z_{GLB} 代表的权重值趋近于 1，否则趋近于 0。全局注意力机制通过全局注意力因子与最后一个 DenseNet 块输出的特征图做元素积运算实现的，表达式如下：

$$H(t)=Y'_{ADB2}(t)*Z_{GLB}(t) \tag{5}$$

其中：$H(t)$ 表示第 t 帧相对于整个音频块的重要程度。因此在音频片段中，全局注意力机制可以通过全局注意力因子 Z_{GLB} 对输出的全局特征加权来选择重要的目标音频事件。另外，在音频事件检测中，输入的音频信号越长意味着输入音频噪声越多，而音频样本中的背景噪声可能导致过拟合问题，通过引入的注意方法可以缓解过拟合问题。

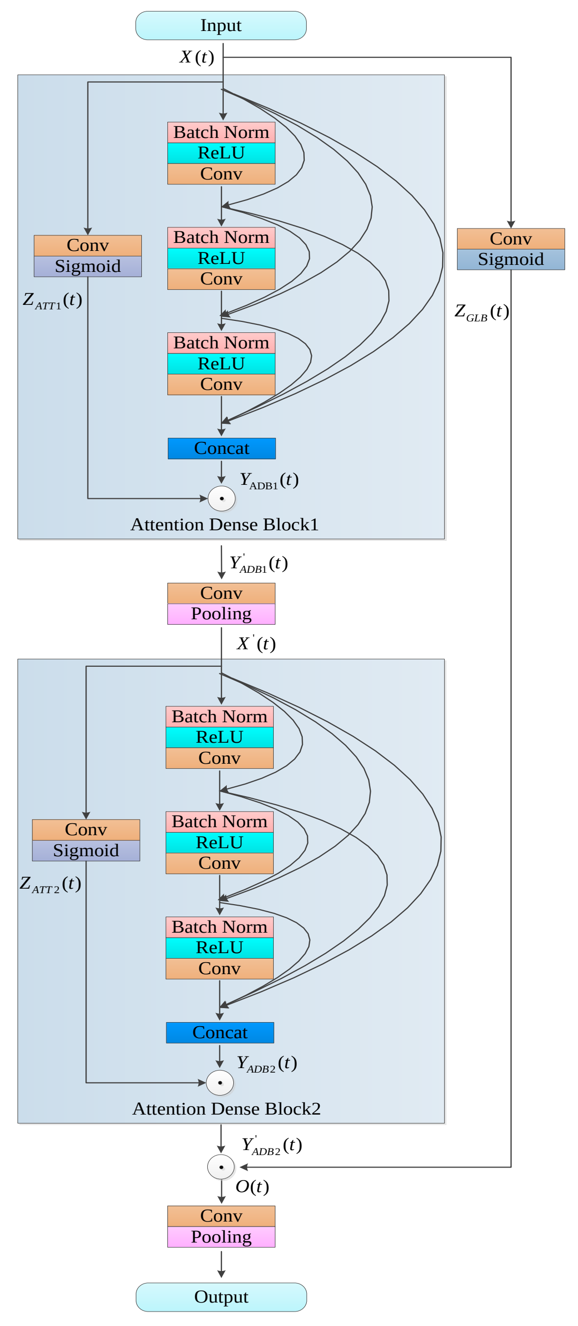


图 2　多层次注意力机制的模块

Fig. 2　The Architecture of the multi-level attention model

1.4　音频事件分类模块

分类层由全连接层、Relu、全连接以及 Sigmoid 输出单元组成。多层次注意力机制层输出的特征传入到前馈层中，前馈层将更新后的特征输入到 Sigmoid 单元来计算目标音频事件出现的概率，即当目标音频事件发生时 Sigmoid 单元输

出值趋近于 1，否则趋近于 0。因此分类层的输出表示目标事件在音频样本中存在的概率。为了对模型进行评估，需要将预测音频事件的概率进行二值化处理，当预测值大于阈值 a_{prob} 时，概率为 1，表示事件存在；否则概率为 0，事件不存在。

2　音频事件检测模型训练算法

音频事件检测模型训练算法如下所示。

算法　音频事件检测模型 MLA-DCNNs 训练算法

输入：原始音频流。

输出：收敛的 MLA-DCNNs 模型。

//构建数据集

1) $\Omega \leftarrow \varnothing$

2) for r = 0 to N-1 do　　　　//N 表示音频流的总数

3)　$X_r = \log-amplitude_mel-spectrogram(audio(r).wav)$ /*提取音频流的对数梅尔谱特征：$X_r=\{x_0,x_1,\cdots,x_{t-1}\}$，其中 t 表示第 t 帧*/

4)　$Y_r = get_label(audio(r).wav)$ /*根据音频流起始和结束时间戳提取对应帧级标签：$Y_r=\{y_0,y_1,\cdots,y_{t-1}\}$，$y_t \in \{0,1\}$，其中数值 1 表示音频事件发生，数值 0 表示音频事件未发生*/

5)　将一个训练实例 $\{X_r, Y_r\}$ 放入 Ω

6) end for

//训练模型

7) 初始化 MLD-DCNNs 模型所有的参数 ϑ_{all}，将数据集随机划分为训练集 Ω_t 和测试集 Ω_v

8) do

9)　从训练集 Ω_t 中随机选取一个 batch 的示例 Ω_b

10)　if(DenseNet 块)　　　//表示一维 DensetNet 层

11)　　for i = 0 to M-1 do　　　//M 表示注意力稠密块的总数

12)　　　if (稠密块)

// L_i 表示第 i 个 ATT-DB 块的卷积层数

13)　　　　for j = 0 to L_i-1 do

14)　　　　　连续 BN-ReLU-Conv(1×1)-BN-ReLU-Conv(3×3)操作

15)　　　　end for

16)　　　end if

17)　　　if(局部注意力块)　　　//计算局部注意力因子

18)　　　　根据公式(2)计算注意力因子 $Z_{ADB(i)}$

19)　　　end if

20)　　　根据公式(3)结合局部注意力因子 $Z_{ADB(i)}$ 计算正向 ATT-DB 块的输出特征 $Y_{ADB}(t)$

21)　　end for

22)　end if　　　　//局部注意力加权结束

23)　if(全局注意力模块)　　　//全局注意力加权

24)　　根据公式(4)(5)结合全局注意力因子 Z_{GLB} 计算正向全局注意力块的输出特征 $H(t)$

25)　end if　　　　//全局注意力加权结束

26)　最后，通过二值交叉熵损失函数求得误差，更新全局参数 ϑ_{all}

27) }while(满足优化条件则停止)

3　实验

3.1　实验数据与实验环境

实验环境：操作系统 windows7，64 位；处理器 Inter(R) Cor-eTMi5-4200M；内存大小为 8 GB；编程平台 Pycharm，Python3.5 版。

实验数据：本次实验采用的数据集来源于 DCASE2017[1] 任务 2 稀疏音频事件检测的开发数据集。数据集包括三类孤立的目标音频事件以及作为背景的日常声学场景。在数据集中，目标音频事件由婴儿哭声、玻璃碎裂声以及枪声组成，背景音频集来自“TUT 声学场景 2016 数据集”的一部分，主要包 15 个不同音频场景。从 freesound.org 上下载三个目标类的孤立事件：婴儿哭声(106 个训练样本，42 个测试样本)，玻璃碎裂声(96 个训练样本，43 个测试样本)，枪声(134 个训练样本，53 个测试样本)与背景音频数据集混合生成 3 000 个混合样本，其中训练集 1 500 个，测试集 1 500 个，在训练集与测试集中每类事件分别有 500 个样本。混合数据集的事件—背景比率为-6 dB、0 dB、6 dB。在每个事件类的 500 个混合样本中，其中一半样本仅含背景音频。

3.2　评价指标

本实验使用了 F1 值以及错误率(error rate，ER)作为评价指标[23]。F1 值与错误率的计算公式如下：

$$F=\frac{2PR}{P+R} \tag{6}$$

$$ER=\frac{FN+FP}{N} \tag{7}$$

其中：P 和 R 分别表示准确率和召回率，其计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{8}$$

$$R=\frac{TP}{TP+FN} \tag{9}$$

其中：TP 表示系统准确的预测了目标事件存在于音频片段中并且成功预测出事件发生的起始位置，成功预测出目标事件的起始时间定义为预测值与实际值误差范围为 500 ms；FP 表示音频样本中不存在目标音频事件，而系统预测目标事件存在；FN 表示音频样本中存在目标音频事件，而系统未能正确预测目标事件存在；N 表示测试数据集中样本总数。

3.3　参数设置

本文使用 keras 和 tensorflow 来搭建模型。在实验中使用了两个注意力稠密模块。ATT-DB 模块的卷积核大小为 3×1，通过调节注意力稠密模块中卷积核的个数以及卷积层的数量来优化网络模型。连接两个注意力稠密模块的转换层模块由 1×1 卷积核以及 2×1 平均池化层组成。转换层模块的压缩因子 θ 设置为 0.5。在训练阶段使用二值交叉熵损失函数作为损失函数，为了优化损失函数，使用 Adam(adaptive momentum)作为优化器，mini-batch 设置为 256。为了防止过拟合，在每一个卷积层后使用 Dropout 层，其值为 0.2。

在分类阶段，预测目标事件是否发生的阈值为 a_{prob}，其范围为 $0 \le a_{prob} \le 1$，采用步长为 0.1 的网格搜索方法获取最优的 F1 值来计算阈值 a_{prob}。如图 3 所示，当阈值 a_{prob} 取 0.7 时，可得到最优的 F1 值，其值为 83.2%。

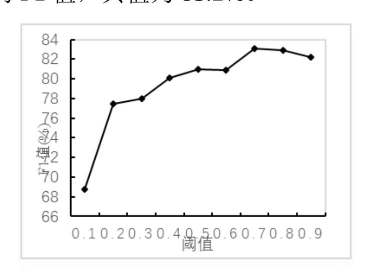


图 3　F1 值随阈值变化趋势

Fig. 3　Trend diagram of F-score changing with threshold value

3.4　实验对比

将本文提出的多层次注意力机制一维 DenseNet 方法 MLA-DCNNs 与以下几种方法进行实验：

a)Baseline[1]。基准系统(DCASE baseline)是 DCASE 2017 挑战赛官方提供的基本方法，主要由含有两个隐藏层的多层感知机模型组成，其中每层隐藏层含有 50 个隐藏单元。

b)CNNs[9]。CNNs 模型是一种端到端的学习模型，能自动从样本数据中学习音频信号的特征，并且有利于提取音频信号的高阶不变特征。

c)DenseNet[12]。DenseNet 模型建立卷积神经网络的前层与后层之间的密集连接。因此，DenseNet 模型增强了特征之间的传递，更有效地利用了特征并缓解了梯度消失问题。

d)CGRNN-Att[17]。 CGRNN-Att 模型在卷积门限循环神经网络基础上引入了注意力机制，其增强了目标音频事件的权重以及抑制不相关的背景音频噪声。

e)CapsNet-Att[18]。 CapsNet-Att 模型在胶囊神经网络模型的基础上引入了注意力机制，主要是由门限卷积网络层、胶囊网络层以及注意力模块组成，详情参见文献[18]。

f)Single-Att[19]。 Single-Att 模型主要是将多示例学习模型与单层注意力机制结合，其主要由三个全连接层和一个注意力模块组成。

3.5　实验结果与实验分析

本文将 MLA-DCNNs 与其他对比方法在 DCASE 2017 任务 2 的开发数据集上用网格搜索的方式进行超参数调优，采用其测试样例进行测试，得到的实验结果如表 1 所示。由表 1 可知，本文提出 MLA-DCNNs 方法的平均 F1 值比基准系统提升了 10.5%，平均错误率 ER 从基准系统 0.53 降低到了 0.31。此外，MLA-DCNNs 方法在婴儿哭声、玻璃破碎声以及枪声等三个类别上 F1 值与错误率 ER 均优于基准系统。由此说明了 MLA-DCNNs 方法的可行性与有效性。

表 1　不同方法的对比

Table 1　Comparison of different method

方法	Baseline		CNNs		DenseNet		GCRNN-Att		CapsNet-Att		Single-Att		MLA-DCNNs	
	ER	F1/%	ER	F1/%	ER	F1/%	ER	F1/%	ER	F1/%	ER	F1/%	ER	F1/%
婴儿哭声	0.67	72	0.42	76.5	0.36	81.2	**0.26**	**88.3**	0.32	82.9	0.32	87.6	0.31	83.5
玻璃破碎声	0.22	88.5	0.19	90.6	0.16	91.3	0.16	91.6	0.17	91.1	0.17	90.4	**0.13**	**93.4**
枪声	0.69	57.4	0.53	67.2	**0.47**	**73.5**	0.54	68.2	0.56	66.7	0.54	65.4	0.49	72.7
平均值	0.53	72.7	0.38	78.1	0.33	82	0.32	82.7	0.35	80.2	0.34	81.1	**0.31**	**83.2**

MLA-DCNNs 方法的 F1 值以及 ER 两个指标均优于 CNNs。原因是 MLA-DCNNs 模型是一种密集网络结构，这使得网络层数更深，可以更好地提取音频信号的高阶不变特征，因此一定程度缓解了背景噪声对音频信号的干扰。

引入多层次注意力机制的 MLA-DCNNs 方法的平均 F1 值比 DenseNet 方法提升了 1.2%，平均错误率 ER 降低了 0.02，这说明引入多层次注意力机制的 DenseNet 网络模型比未引入注意力机制的模型的效果好。此外，在婴儿哭声类中，MLA-DCNNs 相比 DenseNet 方法的错误率有显著改进，原因是婴儿哭声比其他类别音频事件持续时间更长，注意力机制更倾向于关注持续发生的音频事件，因此对持续音频事件婴儿哭声的检测效果更好。在玻璃破碎声类中，MLA-DCNNs 方法在错误率以及 F1 值两个指标上比 DenseNet 方法好，原因是玻璃破碎声与背景音频如街道场景、咖啡馆场景等噪声有显著的区别，而引入注意力机制可以增强目标事件的权重，同时抑制不相关的背景噪声，这说明注意力机制对与背景噪声有显著区别音频事件更有效。

MLA-DCNNs 方法的平均 F1 值比 CGRNN-Att 方法提升了 0.5%，而平均错误率 ER 降低了 0.01。虽然在婴儿哭声类中，MLA-DCNNs 方法在 F1 值和错误率 ER 两个指标上比本文的方法好，但是本文提出的方法的整体检测性能优于 CGRNN-Att 方法。此外，在枪声类中，MLA-DCNNs 方法在 F1 值和错误率 ER 指标上均比 CGRNN-Att 方法好，因为 MLA-DCNNs 方法采用了堆叠的注意力机制，堆叠的注意力机制使得不同层次的注意力感知特性随着网络层数的加深而自适应地变化，因此可以在一定程度上降低枪声产生的混响对音频事件检测造成的影响。

由表 1 可知，本文提出的 MLA-DCNNs 方法整体的性能优于 CapsNet-Att 方法。因为本文采用了一维帧级检测的方法，相比采用二维块级检测 CapsNet-Att 方法，本文提出的帧级检测方法能更有效的检测音频事件的开始和结束时间。

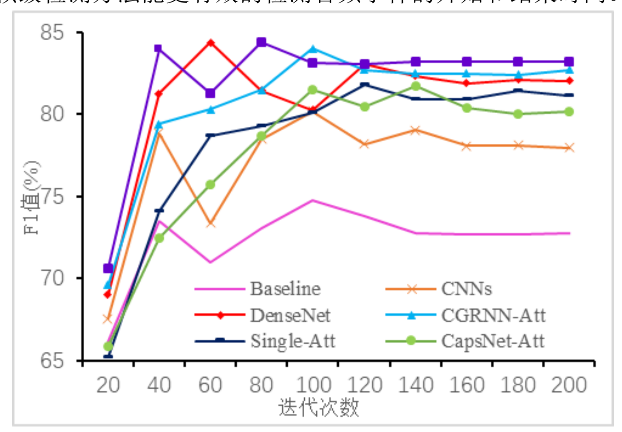


图 4　F1 值迭代变化图

Fig. 4　Diagram of F-score versus number of iterations

相比单层次注意力机制的 Single-Att 方法、MLA-DCNNs 方法的平均 F1 值提升了 2.1%，平均 ER 降低了 0.03，这说明了多层次注意机制比单层次注意力机制能更有效地利用网络的中间层神经元的信息。

图 4 表示随着迭代次数的增加，MLA-DCNNs 模型与其他的对比方法在 DCASE 数据集上 F1 值的变化趋势。由图 4 可以看出，MLA-DCNNs 方法收敛后的 F1 值比其他几种方法高。因此，可以得出引入多层次注意力机制的 MLA-DCNNs 方法有效地提升了模型对音频事件检测任务的建模能力。

4　结束语

本文提出了一种多层次注意机制一维 DenseNet 端到端的网络模型用于音频事件检测。该模型使用一维的 DenseNet 结构可以有效的检测音频事件发生的起始和结束时间，并且通过引入多层次的注意力机制可以使模型关注重要的目标帧以及抑制不相关的背景音频帧来缓解背景噪声对目标音频事件的干扰问题。在 DCASE 2017 任务 2 的开发数据集上的实验结果表明，本文提出的方法的有效性和可行性，对基于深度学习的音频事件检测有一定的贡献。在未来的研究工作中，可以利用多尺度的音频特征作为模型的输入特征对模型进行改进。

参考文献：

[1] Mesaros A, Heittola T, Diment A, *et al.* DCASE 2017 challenge setup: tasks, datasets and baseline system [C]// Proc of DCASE Workshop on Detection and Classification of Acoustic Scenes and Events. 2017.

[2] 屈俊玲，李鸿燕. 基于计算听觉场景分析的混合语音信号分离算法研究 [J]. 计算机应用研究, 2014, 31 (12): 3822-3824. (Qu Junlin, Li Hongyan. Research on separation based on computational auditory

scene analysis [J]. Journal of Computer Application, 2014, 31 (12): 3822-3824.)

[3] Stowell D, Clayton D. Acoustic event detection for multiple overlapping similar sources [C]// Proc of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics . Piscataway, NJ: IEEE Press, 2015: 1-5.

[4] Foggia P, Petkov N, Saggese A, *et al.* Reliable detection of audio events in highly noisy environments [J]. Pattern Recognition Letters, 2015, 65 (C): 22-28.

[5] Wang Yun, Neves L, Metze F. Audio-based multimedia event detection using deep recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2016: 2742-2746.

[6] Mesaros A, Heittola T, Eronen A, *et al.* Acoustic event detection in real life recordings [C]// Proc of the 18th European Signal Processing Conference. Piscataway, NJ: IEEE Press, 2010: 1267-1271.

[7] Mesaros A, Heittola T, Dikmen O, *et al.* Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Brisbane. Piscataway, NJ: IEEE Press, 2015: 151-155.

[8] Phan H, Mazur R, Mertins A. Random regression forests for acoustic event detection and classification [J]. IEEE/ACM Trans on Audio Speech & Language Processing, 2015, 23 (1): 20-31.

[9] Salamon J, Bello J P. Deep convolutional neural networks and data augmentation for environmental sound classification [J]. IEEE Signal Processing Letters, 2017, PP (99): 1-1.

[10] Cakir E, Heittola T, Huttunen H, *et al.* Polyphonic sound event detection using multi label deep neural networks [C]// Proc of International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2015: 1-7.

[11] Dang An, Vu T H, Wang Jiaching. Deep learning for DCASE2017 challenge [R]. DCASE2017 Challenge, Tech. Rep, 2017.

[12] Huang Gao, Liu Zhuang, Maaten L V D, *et al.* Densely connected convolutional networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2261-2269.

[13] Phan H, Krawczyk-Becker M, Gerkmann T, *et al.* Weighted and multi-task loss for rare audio event detection [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2018: 336-340.

[14] Yang Zichao, Yang Diyi, Dyer C, *et al.* Hierarchical attention networks for document classification [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 1480-1489.

[15] Amplayo R K, Kim J, Sung S, *et al.* Cold-start aware user and product attention for sentiment classification [C]// Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2535-2544.

[16] He Ruidan, Lee W S, Ng H T, *et al.* Effective attention modeling for aspect-level sentiment classification [C]// Proc of the 27th International Conference on Computation-al Linguistics. New York: ACM Press, 2018: 1121-1131.

[17] Xu Yong, Kong Qiuqiang, Huang Qiang, *et al.* Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging [C]// Proc of INTERSPEECH. Stockholm, Sweden: ISCA Press, 2017: 3083-3087.

[18] Iqbal T, Xu Yong, Kong Qiuqiang, *et al.* Capsule routing for sound event detection [C]// Proc of the 26th European Signal Processing Conference. Piscataway, NJ: IEEE Press, 2018: 2255-2259.

[19] Kong Qiuqiang, Xu Yong, Wang Wenwu, *et al.* Audio set classification with attention model: a probabilistic perspective [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2018: 316-320.

[20] Lee J, Nam J. Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging [J]. IEEE Signal Processing Letters, 2017, 24 (8): 1208-1212.

[21] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// Proc of the 32th International Conference on Machine Learning. Cambridge, MA: JMLR Press, 2015: 448-456.

[22] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [C]// Proc of the 4th International Conference on Artificial Intelligence and Statistics Piscataway, NJ: IEEE Press, 2011: 315-323.

[23] Mesaros A, Heittola T, Virtanen T. Metrics for polyphonic sound event detection [J]. Applied Sciences, 2016, 6 (6): 162.